学校编码：**10384** 分类号_____密级____

学号：21620100153890 **UDC** ____

厦门大学

博 士 学 位 论 文

一种新的数据不依赖获取质谱数据的分析方法

A novel analysis method for data-independent acquisition MS data

作者姓名：李渊越

指导教师姓名：韩 家 淮 教授

专 业 名 称：生物化学与分子生物学

论文提交日期：2014 年 4 月

论文答辩时间：2014 年 5 月

学位授予日期：2014 年 6 月

答辩委员会主席：_____

评 阅 人：_____

2014 年 **6** 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘要

质谱作为对一种可以对蛋白质进行定性和定量分析的重要工具，一直以来都在生物学研究中受到人们的青睐。传统的数据依赖获取（data-dependent acquisition，DDA）质谱可以有效地对蛋白质进行定性分析，但重复性不高，定量精确度不高。而选择反应监测（Selected Reaction Monitoring，SRM）虽然重复性高、定量准确，但通量过低。近年来新兴的 SWATH 质谱作为一种数据不依赖质谱，很好的解决了上述两种质谱的缺陷，因此受到了人们的广泛关注。但已有的 SWATH 质谱的数据分析方法需要在利用质谱鉴定时往样品中加入 iRT 肽段，此外，在进行分析前，还需要事先构建好肽段特征文库，费时费力，也使得鉴定结果受肽段特征文库含量的限制。

由于在利用质谱解决生物学问题时，通常情况下会有多组样品，且只对多组样品间有差异的蛋白质感兴趣。据此，本文提出了一种新的适用于 SWATH 质谱等数据不依赖获取质谱的数据分析方法，该方法通过比对不同实验组间肽段产生的一级质谱和二级质谱信号强度的相关性，计算出虚拟谱图，再通过数据库搜索的方法鉴定该肽段。该方法很好的解决了以往此类方法中生成虚拟谱图准确性不高的缺点，也不需要事先构建肽段特征文库。利用该方法分析 SWATH-MS 金标准数据库可以鉴定出比现有的分析方法 OpenSWATH 更多的肽段，且具有相当的定量准确性。利用该方法分析肿瘤坏死因子刺激下的 L929 细胞可以鉴定出已知的蛋白相互作用，并能发现一些未知的蛋白相互作用。

关键词：数据不依赖质谱；分析方法

Abstract

As an important tool for qualitative and quantitative analysis of proteins, mass spectrum (MS) has been popular in biological research since its emergence. Traditional data-dependent acquisition (DDA) MS can effectively operate qualitative analysis of proteins, but its repetitiveness and quantitative function are weak. Another method, selected reaction monitoring (SRM), though with high repetitiveness and quantitative accuracy, has been limited by its low flux. The recently developed SWATH MS can make up the defects of the two MS methods above. And as a date-independent MS technology, it draws the extensive attention of people. But current data analysis method of SWATH MS requires the addition of iRT peptide to the prepared samples before the identification. On the other hand, a characterized peptide library should be constructed before the analysis, which requires much time and energy, also limits the results with its capacity. Normally, when solving biological issues with MS, the researches must deal with multiple samples to find the differences in proteins. This thesis put forward a novel analysis method which could be used for analyzing data from data-independent acquisition MS, including SWATH MS. This method can identify peptides by calculating the virtual spectrum through comparison of the relativity between a mass spectrometry and two-stage mass spectrometry, which comes from peptides of different groups, and searching the database. It has higher accuracy of the virtual spectrum than previous methods, while has no need of a characterized peptide library. Furthermore, it can identify more peptides than OpenSWATH when analyzing the SWATH-MS standard database, with commendable quantitative accuracy. We can use this novel analysis method to identify previously reported protein-protein interactions in tumor

necrosis factor (TNF)-stimulated L929 cells, with revealing of some unknown interactions.

Keywords: data-independent acquisition mass spectrum; method for analysis

目录

第一章 前言 1
1.1 质谱 1
1.1.1 常见的用于蛋白质组学研究的质谱 1
1.1.2 利用质谱分析蛋白组的基本策略 2
1.1.2.1 自下而上的研究策略 3
1.1.2.2 中间向下的研究策略 4
1.1.2.3 自上而下的研究策略 4
1.1.3 利用质谱分析蛋白组的基本流程 4
1.2 质谱的采集方式 5
1.2.1 信息依赖的采集方式 5
1.2.1.1 鸟枪法（Shotgun）采集方式 5
1.2.1.2 定向（Directed）采集方式 6
1.2.1.3 目标法（Target）采集方式 7
1.2.2 信息不依赖（Data Independent Acquisition，DIA）的采集方式 8
1.2.2.1 不选择母离子的采集方式 8
1.2.2.1.1 离子化时解离技术 8
1.2.2.1.2 UPLC/MS^{E}技术 9
1.2.2.1.3 所有离子解离（all-ion fragmentation，AIF）技术 10
1.2.2.2 选择母离子的采集方式 11
1.2.2.2.1 傅立叶变化-所有反应监测（Fourier transform-all reaction monitoring，FT-ARM）技术 11
1.2.2.2.2 不依赖离子数目的母离子获取（precursor acquisition independent from ion count, PAcIFIC）技术 11
1.2.2.2.3 UDMS^{E}技术 12

1.2.2.2.4 SWATH技术 13
1.3 质谱定量技术 14
1.3.1 代谢标记技术 15
1.3.2 蛋白标记技术 16
1.3.3 肽段标记技术 16
1.3.4 非标记定量技术 17
1.3.4.1 基于谱图计数的定量 17
1.3.4.2 基于一级质谱强度的定量 17
1.3.4.3 选择反应监测（Selected reaction monitoring, SRM）技术 17
1.4 数据分析技术 19
1.4.1 峰的提取 19
1.4.2 质谱时间对齐 19
1.4.2.1 基于原始数据的对齐方法 20
1.4.2.1.1 基于总离子电流的对齐 20
1.4.2.1.2 基于具体谱图的对齐 20
1.4.2.2 基于特征的对齐方法 21
1.4.2.2.1 基于色谱峰的对齐 21
1.4.2.2.2 基于图像的对齐 21
1.4.2.2.3 基于显著特征的对齐 21
1.4.2.3 驻留时间预测 21
1.4.2.3.1 基于算法的驻留时间预测 21
1.4.2.3.2 基于标准样品的驻留时间预测 22
1.5 信息不依赖采集方式的分析方法 22
1.5.1 谱图计算的方法 23
1.5.1.1 一级质谱不依赖的方法 23
1.5.1.2 一级质谱依赖的方法 23
1.5.2 文库依赖的方法 25
第二章 材料和方法 27

2.1 常用药品和试剂 27
2.2 实验仪器 27
2.3 DNA相关实验方法 28
2.3.1 质粒载体 28
2.3.1.1 pBOBI 28
2.3.1.2 pTK-Neo-USER-3Flag 29
2.3.2 大肠杆菌感受态细胞的制备 30
2.3.3 DNA转化 31
2.3.4 质粒DNA的提取 31
2.3.4.1 小规模质粒DNA的提取（STET煮沸法） 31
2.3.4.2 中等规模质粒DNA的提取（碱变性法） 32
2.3.4.3 大规模质粒DNA的提取（氯化铯密度梯度离心法） 33
2.3.5 质粒DNA的工具酶处理 34
2.3.5.1 DNA的限制性内切酶消化 34
2.3.5.2 线性DNA 5’端磷酸基团的去除 34
2.3.6 DNA扩增反应（PCR反应） 35
2.3.7 DNA片段的纯化 35
2.3.7.1 琼脂糖电泳分离DNA 35
2.3.7.2 DNA的回收 36
2.3.8 DNA连接反应 36
2.3.9 Libarary Independent Clone（LIC） 36
2.3.10 哺乳动物细胞表达质粒的构建 37
2.3.11 原核表达质粒的构建 37
2.4 细胞相关实验 37
2.4.1 细胞培养 37
2.4.1.1 细胞培养液的配制 37
2.4.1.2 细胞的传代和接种 38
2.4.2 瞬时转染 38
2.4.2.1 Lipofectamine 2000 转染 38

2.4.2.2 磷酸钙转染 38
2.4.3 慢病毒感染 39
2.4.4 利用基因敲入（knock in）构建稳定表达细胞系 39
2.4.4.1 克隆的构建 39
2.4.4.2 病毒的收集 40
2.4.4.3 AAV病毒感染L929细胞和单克隆筛选 40
2.4.4.4 基因组DNA的提取 41
2.4.4.5 PCR鉴定阳性克隆 41
2.4.4.6 Neo筛选标记的切除 41
2.5 蛋白质相关实验方法 42
2.5.1 免疫共沉淀 42
2.5.2 大肠杆菌中融合蛋白的表达和纯化 43
2.5.3 琼脂糖珠吸附融合蛋白的洗脱 44
2.5.4 蛋白浓度测定 44
2.5.5 肽段样品制备 45
2.5.6 酶解肽段的C18反相色谱柱脱盐 46
2.6 数据处理相关实验方法 47
2.6.1 数据来源 47
2.6.2 数据库搜索 47
第三章 结果与讨论 48
3.1 分析原理 48
3.1.1 一级质谱与二级质谱的信号强度相关性的分析 48
3.1.2 不依赖于文库的SWATH数据分析的基本原理 49
3.2 分析流程 52
3.2.1 LISWATH的分析流程 52
3.2.2 LISWATH的依赖关系 54
3.3 算法描述 54
3.3.1 LISWATH的基本算法 54
3.3.2 预处理 56

3.3.3 驻留时间对齐 56
3.4 对金标准样品的分析结果 57
3.4.1 LISWATH对驻留时间进行校正的结果分析 57
3.4.2 LISWATH对肽段鉴定结果的分析 60
3.4.3 LISWATH对肽段鉴定结果可信度的分析 63
3.4.4 LISWATH定量精确度的分析 65
3.5 对真实样品的分析结果 67
3.5.1 利用LISWATH分析Flag-TNF的免疫沉淀样品 68
3.5.2 利用LISWATH分析Flag-RIPK1 的免疫沉淀样品 69
3.5.3 利用LISWATH分析Flag- RIPK3 的免疫沉淀样品 70
3.6 讨论 71
3.6.1 不同质谱获取方式的比较 71
3.6.2 现有的数据不依赖获取质谱分析方法优缺点的比较 73
参考文献 75
致谢 81
附录 I 图表索引 82

Table of Contents

CHAPTER 1 INTRODUCTION.....1
1.1 Mass Spectrum (MS).....1
1.1.1 Common MS in proteomics study.....1
1.1.2 Basic strategy for proteomics study using MS.....2
1.1.2.1 The bottom-up strategy.....3
1.1.2.2 The middle-down strategy.....4
1.1.2.3 The top-down strategy.....4
1.1.3 The basic flow of proteomic analysis using MS.....4
1.2 Acquisition method of MS.....5
1.2.1 Data-dependent acquisition.....5
1.2.1.1 Shotgun acquisition.....5
1.2.1.2 Directed acquisition.....6
1.2.1.3 Target acquisition.....7
1.2.2 Data-independent acquisition.....8
1.2.2.1 Precursor non-selective acquisition.....8
1.2.2.1.1 Ionization-dissociation technology.....8
1.2.2.1.2 UPLC/MS^E technology.....9
1.2.2.1.3 UPLC/MS^E technology.....10
1.2.2.2 Precursor-selective acquisition.....11
1.2.2.2.1 Fourier transform-all reaction monitoring technology.....11
1.2.2.2.2 Precursor acquisition independent from ion count technology.....11
1.2.2.2.3 $UDMS^E$ technology.....12
1.2.2.2.4 SWATH technology.....13
1.3 Quantitative MS.....14

1.3.1 Metabolic labeling.....15
1.3.2 Protein labeling.....16
1.3.3 Peptide labeling.....16
1.3.4 Label free quantitative technology.....17
1.3.4.1 Quantization based on spectrum amounts.....17
1.3.4.2 Quantization based on spectrum amounts.....17
1.3.4.3 Selected reaction monitoring technology.....17
1.4 MS data analysis.....19
1.4.1 Extraction of the spectrum peaks.....19
1.4.2 Time alignment.....19
1.4.2.1 Alignment based on original data.....20
1.4.2.1.1 Alignment based on total ion current.....20
1.4.2.1.2 Alignment based on particular spectrum.....20
1.4.2.2 Alignment based on feature.....21
1.4.2.2.1 Alignment based on chromatographic peak.....21
1.4.2.2.2 Alignment based on images.....21
1.4.2.2.3 Alignment based on distinct characterization.21
1.4.2.3 Prediction of retent time.....21
1.4.2.3.1 Prediction of retent time based on algorithm.21
1.4.2.3.2 Prediction of residence time based on iRT....22
1.5 Analysis methods of data-independent acquisition MS data.22
1.5.1 Spectrum calculation-based method.....23
1.5.1.1 Mass spectrum 1 independent method.....23
1.5.1.2 Mass spectrum 1 dependent method.....23
1.5.2 Library dependent method.....25
CHAPTER 2 MATERIALS AND METHODS.....27
2.1 Drugs and reagents.....27
2.2 Experimental equipments.....27
2.3 DNA-related protocols.....28

2.3.1 Vectors.....28
2.3.1.1 pBOBI.....28
2.3.1.2 pTK-Neo-USER-3Flag.....29
2.3.2 Preparation of E.coli competent cells.....30
2.3.3 DNA transformation.....31
2.3.4 DNA purification.....31
2.3.4.1 Mini-scale DNA purification (STET boil).....31
2.3.4.2 Medium-scale DNA purification(Alkaline denaturation).....32
2.3.4.3 Large-scale DNA purification (CsCl gradient centrifugation).....33
2.3.5 Enzymatic manipulation of plasmid DNA.....34
2.3.5.1 Restriction enzyme digestion of DNA.....34
2.3.5.2 Removal of 5'-phosphate group of linear DNA.....34
2.3.6 DNA amplification (PCR reaction).....35
2.3.7 DNA amplification (PCR reaction).....35
2.3.7.1 DNA separation with agarose electrophoresis.....35
2.3.7.2 DNA extraction.....36
2.3.8 DNA ligation.....36
2.3.9 Libarary Independent Clone (LIC).....36
2.3.10 Libarary Independent Clone (LIC).....37
2.3.11 Construction of prokaryotic-expressing plasmid.....37
2.4 Cell-related protocols.....37
2.4.1 Cell culture.....37
2.4.1.1 Preparation of cell culture medium.....37
2.4.1.2 Preparation of cell culture medium.....38
2.4.2 Transient transfection.....38
2.4.2.1 Transfection with Lipofectamine 2000.....38
2.4.2.2 Transfection with calcium phosphate.....38

2.4.3 Lenti-virus infection.....39
2.4.4 Construction of stable expression cell line through gene knock in.....39
2.4.4.1 Construction of plasmid.....39
2.4.4.2 Collection of virus.....40
2.4.4.3 AAV infect L929 cells and single clone selection.40
2.4.4.4 Preparation of genomic DNA.....41
2.4.4.5 PCR identify positive clone.....41
2.4.4.6 Excision of Neo selection marker.....41
2.5 Protein-related protocols.....42
2.5.1 Immuno co-precipitation.....42
2.5.2 Expression and purification of fusion proteins from E.coli 43
2.5.3 Elution of beads-adsorbed fusion proteins.....44
2.5.4 Measurement of protein concentration.....44
2.5.5 Preparation of peptide samples.....45
2.5.6 Desalting of enzyme-digested peptides with C18 reversed-phase chromatographic column.....46
2.6 Data processing-related protocols.....47
2.6.1 Data source.....47
2.6.2 Database retrieval.....47
CHAPTER 3 RESULTS AND DISCUSSION.....48
3.1 Principles of analysis.....48
3.1.1 Relativity analysis between a mass spectrometry and two-stage mass spectrometry.....48
3.1.2 Principles of library-independent SWATH data analysis49
3.2 Workflow of analysis.....52
3.2.1 Workflow of analysis with LISWATH.....52
3.2.2 The dependency of LISWATH.....54

3.3 Algorithm description.....54
3.3.1 algorithm of LISWATH.....54
3.3.2 Prescan.....56
3.3.3 Alignment of retent time.....56
3.4 Analysis of retent time correction with LISWATH.....57
3.4.1 Analysis of retent time correction with LISWATH.....57
3.4.2 Analysis of peptide identification with LISWATH.....60
3.4.3 Analysis of reliability of peptide identification with LISWATH.....63
3.4.4 Analysis of LISWATH quantitative accuracy.....65
3.5 Analysis results of real samples.....67
3.5.1 Analysis of Flag-TNF IP samples with LISWATH.....68
3.5.2 Analysis of Flag-RIPK1 IP samples with LISWATH.....69
3.5.3 Analysis of Flag- RIPK3 IP samples with LISWATH.....70
3.6 Discussion.....71
3.6.1 Comparison of different MS data acquisition methods..71
3.6.2 Comparison of current data-independent acquisition MS73
REFERENCES.....75
ACKNOWLEDGE 81
ADDENDUM I Index of figures and tables.....82